*Nucleic Acids Research, 2004, Vol. 32, Database issue* ***D185–D188***
*DOI: 10.1093/nar/gkh002*

# ArchDB: automated protein loop classification as a tool for structural genomics

**Jordi Espadaler[1,2], Narcis Fernandez-Fuentes[1,3], Antonio Hermoso[1], Enrique Querol[1], Francesc X. Aviles[1], Michael J. E. Sternberg[3] and Baldomero Oliva[2,*]**

[1]Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain, [2]Laboratori de Bioinformàtica Estructural, Grup de Recerca d'Informàtica Biomédica—IMIM, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, C/Doctor Aiguader 80, Barcelona 08003, Catalonia, Spain and [3]Structural Bioinformatics Group, Biochemistry Building, Department of Biological Sciences, Imperial College, London SW7 2AZ, UK

Received March 28, 2003; Revised and Accepted May 2, 2003

## ABSTRACT

**The annotation of protein function has become a crucial problem with the advent of sequence and structural genomics initiatives. A large body of evidence suggests that protein structural information is frequently encoded in local sequences, and that folds are mainly made up of a number of simple local units of super-secondary structural motifs, consisting of a few secondary structures and their connecting loops. Moreover, protein loops play an important role in protein function. Here we present ArchDB, a classification database of structural motifs, consisting of one loop plus its bracing secondary structures. ArchDB currently contains 12 665 super-secondary elements classified into 1496 motif subclasses. The database provides an easy way to retrieve functional information from protein structures sharing a common motif, to search motifs found in a given SCOP family, superfamily or fold, or to search by keywords on proteins with classified loops. The ArchDB database of loops is located at http://sbi.imim.es/archdb.**

## INTRODUCTION

Loops are regions of non-repetitive conformation connecting regular secondary structures. There have been many attempts to classify loops, presenting topological clusters and consensus sequences (1–6). The reports of Salem *et al.* (7) and Wood and Pearson (8) suggested that folds are mainly made up of a number of simple local units of super-secondary structural motifs, formed by a few secondary structures connected by loops. An elementary super-secondary motif can be defined as one loop plus its bracing secondary structures. In particular, loops play an important role in the local conformation (9) of the protein and are often related to its function.

Structural genomics initiatives attempt to infer details of protein function via 3D structure determination (10,11). If a new protein structure adopts a previously observed fold, then functional details might be inferred by considering the function of other proteins adopting the same fold (12–15). If fold similarities are ambiguous or if a protein adopts a new fold, it is still possible to infer function by comparison of key active site residues (16,17). Common detected structural motifs contain particularly useful information on the conservation of specific residues across species, being occasionally involved in the protein function (i.e. the activation loop of some kinases) or in the folding nucleus (18). Moreover, loops are often the most difficult structures to model (6,19) and thus a database of structurally classified protein loops will have widespread applications (i.e. in model building or to complete locally undefined regions from an X-ray diffraction map).

In a previous publication (4), we presented a semi-automated classification of protein loops from a non-redundant database of proteins (20), based on loop conformation and bracing secondary structure type and geometry. However, little or no categorization was obtained for the majority of long loops. Now, we have updated and fully automated the clustering of protein loops and also obtained clusters for many long loops. ArchDB is a web based classification of structural motifs consisting of segments of one loop plus the bracing secondary structures.

## IMPROVED LOOP CLASSIFICATION PROTOCOL

The current version of ArchDB is based on a list of protein domains derived from SCOP 40 on release 1.61 of SCOP (12). Structures not obtained by X-ray crystallography or with resolution greater than 3.0 Å were removed, resulting in a list of 2458 chains (3616 SCOP domains), from which 39 330 motifs (loops plus their bracing secondary structures) were extracted.

We have used the loop clustering program Arch-Type to derive a fully automated loop classification of clusters with more than two loops. The algorithm clustering is based on a density search on the ($\phi$,$\psi$) space of the loop conformation, henceforth allowing for a second check by RMSD. Clusters were arranged as in the previous work. In the lowest level of the classification, structural motifs were grouped according to their geometry (motif subclass level). At higher levels, motifs

*To whom correspondence should be addressed. Tel: +34 93 2240880; Fax: +34 93 2240875; Email: boliva@imim.es

 

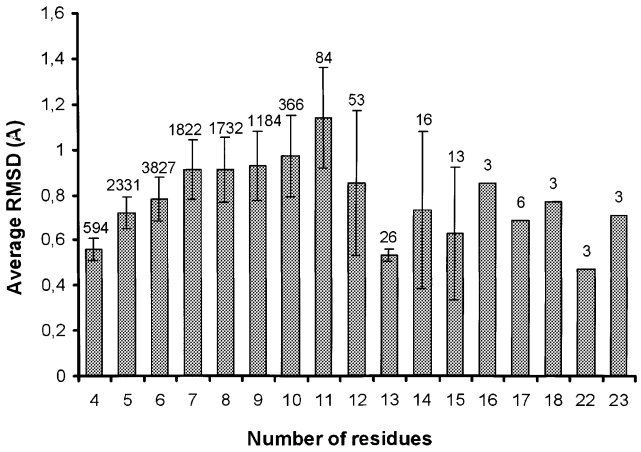


**Figure 1.** Averaged RMSD between loops in subclasses. The averaged RMSD of the sets of loop structures on each subclass was calculated with the main-chain atoms of the residues in the loop plus two bracing residues at each side. Additional extensions of the bars show the standard deviations of the averages with the total loops involved in the RMSD calculation shown at the top. Due to the dramatic decrease in loops with length larger than 11 residues, the significance of the average RMSD is also waning.

were grouped according to the loop size and ($\phi$,$\psi$) conformation (motif class level). At the top of the classification, motifs were identified according to bracing secondary structure type ($\alpha$–$\alpha$, $\beta$–$\beta$ links, $\beta$–$\beta$ hairpins, $\alpha$–$\beta$ and $\beta$–$\alpha$, the motif type level). At the class level, loops of similar size, with differences of $\pm 1$ residue, were allowed to cluster together to deal with the lax definition of the secondary structure ends. Owing to the $\pm 1$ extension and to the wide definition around ($\phi$,$\psi$) regions in l/g and in b/p conformations, loops were allowed to cluster into more than one group. A reclustering protocol has been devised to deal with the overlap between clusters. Overlapping clusters are merged depending on the percentage of shared loops (see Supplementary Material). The result is an optimized partition of the conformational space of loops that groups clusters [as obtained in Arch-Type (4) and containing at least two loops] into subclasses with the largest number of loops and the minimum overlap. Finally, the averaged RMSD between loops in each subclass was checked in order to corroborate this procedure (Fig. 1). Each subclass is identified in ArchDB by a three-number code as defined in the original paper (4).

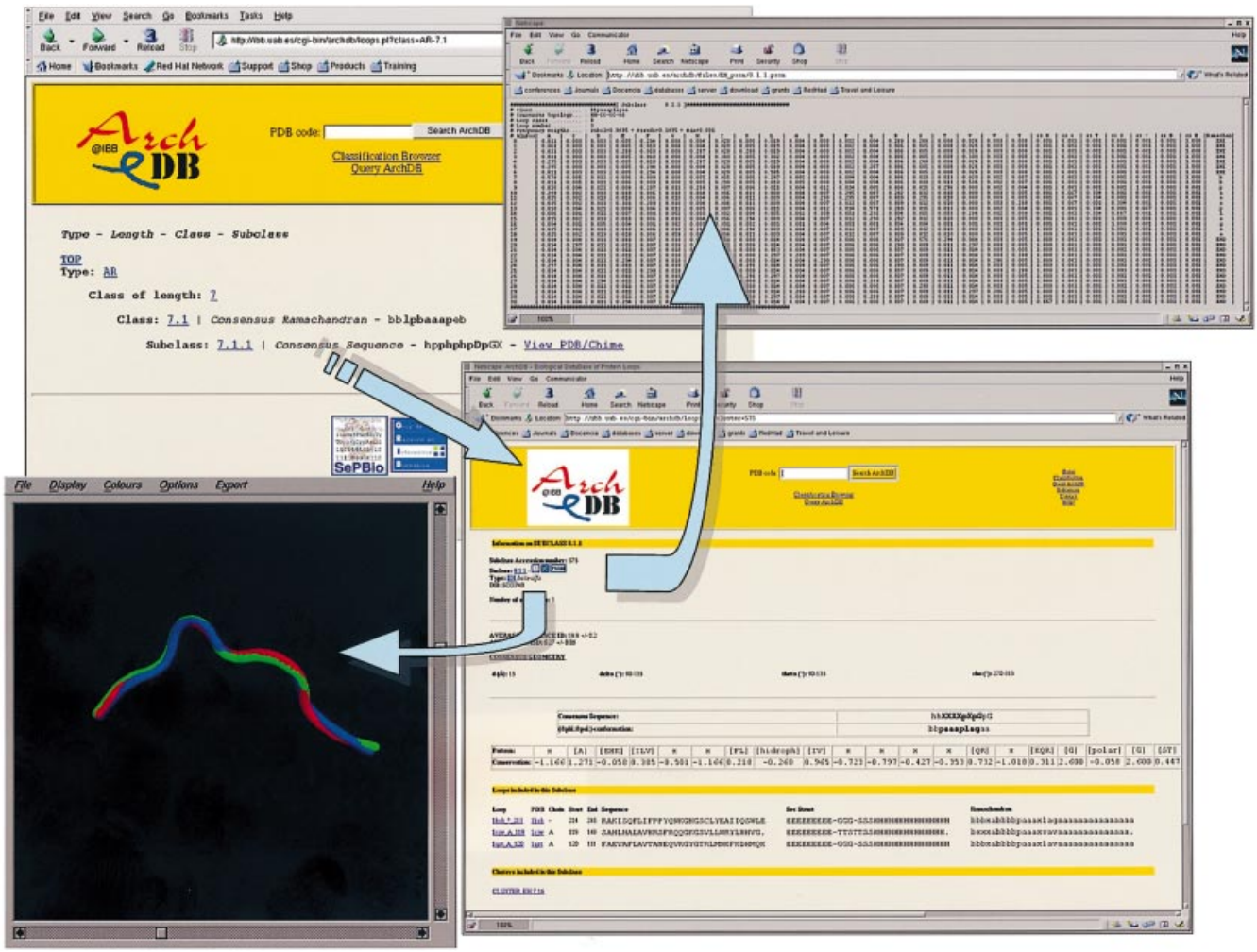


**Figure 2.** Screenshot of ArchDB information HTML pages. Classification browser with subclass information, multiple alignments of sequence and conformation, the profile pattern and the image of no more than four structurally superimposed motifs as viewed with Rasmol (27).

**Table 1.** Total of classes, subclasses and loops classified

| Loop type | Classes | Subclasses | Loops |
|---|---|---|---|
| β-link | 96 | 218 | 1219 |
| β-hairpins | 76 | 222 | 3150 |
| α–α | 70 | 249 | 1968 |
| α–β | 108 | 443 | 3734 |
| β–α | 101 | 364 | 2594 |

Shows the total of loops, classes and subclasses as obtained for the SCOP40 v1.61 based classification.

After comparing the classifications of loops obtained with the PDB (21) structures of PDB_SELECT (20) at 25% from years 2000 to 2002, and databases obtained with PDB_SELECT at 25% and 35% and SCOP40 from release1.61 of SCOP, we found that SCOP40 yielded a more stable classification (classes and subclasses are more conserved between updates; see Supplementary Material).

## FEATURES

Users can query the ArchDB database by six different methods: (i) search for structure motifs found within a PDB structure by specifying the PDB identifier (21) or SWISS-PROT (22) accession code; (ii) browse through ArchDB classes and subclasses; (iii) retrieve structure motifs satisfying some features [i.e. bracing secondary structure type, loop size and loop (ϕ,ψ) conformation]; (iv) search for structural motifs found within a SCOP family/superfamily/fold; (v) search for SWISS-PROT keywords or GO accession codes; (vi) search for motifs simultaneously found in two different PDB structures (regardless of the fold type).

A table describing the consensus (more than 80% common sequence/conformation), geometry, and loop membership—identified by PDB code, chain ('*' for null) and first residue—is displayed for each subclass. Additional information includes the average percentage of sequence identity and averaged RMSD of main-chain atoms. Also a PROSITE-like pattern (23), with the position-specific entropy as calculated with the program AL2CO (24), and the PSSM profile derived from the sequence multiple alignment are included. 3D images of superimposed super-secondary motifs can be viewed using Rasmol or Chime. Multiple alignment of sequences, secondary structures and (ϕ,ψ) conformations are provided (Fig. 2). Structural and functional information for each structure are accessible, including resolution, R-factor, PDB source, SWISS-PROT keywords, GeneOntology (25) or/and Enzyme (22) annotation, and the SCOP domain classification. ArchDB includes links to these databases.

## CURRENT DATABASE CONTENT

A total of 12 665 super-secondary structures out of 39 330 (as found on the starting dataset) were clustered. ArchDB currently contains 451 motif classes and 1496 motif subclasses (Table 1). Each subclass contains a minimum of two loops. After applying the reclustering protocol, the average overlap was 0.9% between motif classes and 0.5% between motif subclasses.

A total of 582 folds out of 701, and 1548 families out of 1940, from SCOP v1.61 have a representative in ArchDB. Folds not in SCOP40 do not have representatives in ArchDB (in this current database). Also, proteins with non-regular secondary structure (NORs) (26) cannot be found in this classification because of the intrinsic definition of loop. Some well-known functional motifs were found among classified loops in ArchDB (Table 2) split by geometry and conformation, as for example: six subclasses with different geometries containing the P-loop, two different subclasses containing the NAD-binding motif, three subclasses containing the EF-hand motif, three subclasses with loops from kinases (two for the catalytic loop, HRD-loop, and one for the activation loop, DFG-loop), or a canonical loop (L1) from immunoglobulins.

## FUTURE DIRECTIONS

ArchDB functional information will be expanded by including ligand information from PDB structures, by developing a protocol to allow the automated functional annotation of structural motifs, and by including new starting protein sets of known structure (i.e. enzymes with known 3D structure, SCOP90, etc.).

**Table 2.** Examples of known functional motifs found in ArchDB

| Subclass | Functional motif | (ϕ,ψ) conformation | Sequence consensus |
|---|---|---|---|
| AR 5.9.1 | Kinases: activation loop (DFG) | bb{***laaap***}bb | hh{***DFGhh***}Xh |
| AR 6.2.1 | $H_1$ canonical loop | bb{***lbpaa***}bp | hS{***GhpFpp***}YW |
| AR 8.2.1 | Kinases: catalytic loop (HRD) | bb{***alapbaaa***}bb | Ih{***HXDhpPpN***}hh |
| AR 8.2.2 | Kinases: catalytic loop (HRD) | bb{***alapbaaa***}bb | hh{***HRDLphXN***}hL |
| EH 3.1.1 and EH 3.1.2 | NAD(P)-binding loop | bb{***eap***}aa | hh{***GXG***}Xh |
| EH 6.1.1 and EH 6.1.2 | P-loop | bb{***eppgag***}aa | hh{***GXXGXG***}Kp |
| EH 6.1.3 | P-loop | bb{***eppgag***}aa | hh{***GhhXXG***}Kp |
| EH 6.4.1 | P-loop | bb{***epplag***}aa | hh{***GpXpXG***}Kp |
| EH 6.4.2 | P-loop | bb{***epplag***}aa | hX{***GXXXXG***}Kp |
| HE 6.6.1 | EF-hand | aa{***paalal***}bb | ph{***DhDpDG***}pI |
| HE 6.6.2 | EF-hand | aa{***paalal***}bb | hh{***DTNXDG***}ph |
| HH 9.1.1 | EF-hand | aa{***paalalbbp***}aa | ph{***DXDXpGXhp***}Xp |

Subclasses were identified by motif type and the three number code (see text). Motif names shown in the table were extracted from the literature. Regions in (ϕ,ψ) space are α, $\alpha_L$, γ, β, $\beta_P$ and ε (described here as **a**, **l**, **g**, **b**, **p** and **e**). Consensus sequence was derived from the multiply aligned loops. A code was used to represent the chemical properties of the consensus at every position: (1) the one letter amino acid code; (2) **p** for polar; (3) **h** for non-polar residues; (4) **X** denotes no sequence consensus.

## SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

## ACKNOWLEDGEMENTS

Supporting grants from Fundacion Areces and from MCYT (Ministerio de Ciencia y Tecnologia, Spain; ref. BIO2002-03609) are acknowledged by B.O. Also, from MCYT (ref. BIO2001-246 and BIO2001-264) and CERBA (Centre de Referencia en Biotecnologia, Generalitat de Catalunya) by F.X.A and E.Q. Support from predoctoral fellowships from the Generalitat de Catalunya and MCYT (Spain) are acknowledged by J.E. and N.F.F.

## REFERENCES

1. Efimov,A.V. (1993) Patterns of loop regions in proteins. *Curr. Opin. Struct. Biol.*, **3**, 379–384.
2. Wintjens,R.T., Rooman,M.J. and Wodak,S.J. (1996) Automatic classification and analysis of αα-turn motifs in proteins. *J. Mol. Biol.*, **255**, 235–253.
3. Kwasigroch,J.M., Chomilier,J. and Mornon,J.P. (1996) A global taxonomy of loops in globular proteins. *J. Mol. Biol.*, **259**, 855–872.
4. Oliva,B., Bates,P.A., Querol,E., Avilés,F.X. and Sternberg,M.J. (1997) An automatic classification of the structure of protein loops. *J. Mol. Biol.*, **266**, 814–830.
5. Oliva,B., Bates,P.A., Querol,E., Avilés,F.X. and Sternberg,M.J. (1998) Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction. *J. Mol. Biol.*, **279**, 1193–1210.
6. Burke,D.F., Deane,C.M. and Blundell,T.L. (2000) Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. *Bioinformatics*, **16**, 513–519.
7. Salem,G.M., Hutchinson,E.G., Orengo,C.A. and Thornton,J.M. (1999) Correlation of observed fold frequency with the occurrence of local structural motifs. *J. Mol. Biol.*, **287**, 969–981.
8. Wood,T.C. and Pearson,W.R. (1999) Evolution of protein sequences and structures. *J. Mol. Biol.*, **291**, 997–995.
9. Yang,A.S. and Wang,L.Y. (2002) Local structure-based sequence profile database for local and global protein structure predictions. *Bioinformatics*, **18**, 1650–1657.
10. Eisenberg,D., Marcotte,E.M., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.
11. Shapiro,L. and Harris,T. (2000) Finding function through structural genomics. *Curr. Opin. Biotechnol.*, **11**, 31–35.
12. Murzin,A.G. (1996) Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.*, **2**, 895–903.
13. Russell,R.B., Saqi,M.A., Sayle,R.A., Bates,P.A. and Sternberg,M.J. (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.*, **269**, 423–439.
14. Dietmann,S., Park,J., Notredame,C., Heger,A., Lappe,M. and Holm,L. (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.*, **29**, 55–57.
15. Dietmann,S., Fernandez-Fuentes,N. and Holm,L. (2002) Automated detection of remote homology. *Curr. Opin. Struct. Biol.*, **12**, 362–367.
16. Russell,R.B., Sasieni,P.D. and Sternberg,M.J. (1998) Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.*, **282**, 903–918.
17. Hegyi,H. and Gerstein,M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.*, **288**, 147–164.
18. Mirny,L. and Shakhnovich,E. (2001) Evolutionary conservation of the folding nucleus. *J. Mol. Biol.*, **308**, 123–129.
19. Fiser,A., Do,R.K. and Sali,A. (2000) Modelling of loops in protein structures. *Protein Sci.*, **9**, 1753–1773.
20. Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Sci.*, **1**, 409–417.
21. Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliard,G.L., Bluhm,W., Weissig,H., Greer,D.S. *et al.* (2002). The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.
22. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
23. Falquet,L. Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
24. Pei,J. and Grishin,N. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.
25. Ashburner,M., Ball,C.A. Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski, K., Dwigth, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
26. Liu,J., Tan,H. and Rost,B. (2002). Loopy proteins appear conserved in evolution. *J. Mol. Biol.*, **322**, 53–64.
27. Sayler,R. and Millner-White,E. (1995). RASMOL: Biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.